Discrete associated kernels method and extensions

Célestin C. Kokonendji [a] and

[a] *Université de Franche-Comté, UFR Sciences et Techniques*
Laboratoire de Mathématiques de Besançon - UMR 6623 CNRS
16 route de Gray – 25030 Besançon cedex, France

Tristan Senga Kiessé [b,*]

[b] *Office National des Forêts - Centre de Nancy*
54840 Velaine-en-Haye, France

Abstract

Discrete kernel estimation of a probability mass function (p.m.f.) often mentioned in the literature has been far less investigated in comparison with continuous kernel estimation of a probability density function (p.d.f.). In this paper, we are concerned with a general methodology of discrete kernels for smoothing a p.m.f. f. We give a basic of mathematical tools for further investigations. First, we point out a generalizable notion of discrete associated kernel which is defined at each point of the support of f and built from any parametric discrete probability distribution. Then, some properties of the corresponding estimators are shown, in particular pointwise and global (asymptotical) properties. Other discrete kernels are constructed from usual discrete probability distributions such Poisson, binomial and negative binomial. For small samples sizes, underdispersed discrete kernel estimators are more interesting than the empirical estimator; thus, an importance of discrete kernels is illustrated. The choice of smoothing bandwidth is classically investigated according to cross-validation and, novelly, to excess of zeros methods. Finally, a unification way of this method concerning general probability function is discussed.

Key words: Asymmetric kernel, dispersion, cross-validation, discrete distribution, finite difference, mean integrated squared error, nonparametric estimation, variable kernel, zero-proportion.
MSC 2010: Primary 62G07 ; Secondary 62G99.

* *Corresponding author*
Email addresses: celestin.kokonendji@univ-fcomte.fr (Célestin C. Kokonendji), tristan.senga-kiesse@onf.fr (Tristan Senga Kiessé).

1 Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with an unknown probability density function (p.d.f.) f on $\mathbb{R}$. A continuous kernel estimator $\widetilde{f}_n$ of f can be defined in the two following ways:

$$\widetilde{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

$$= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}\left(X_i\right), \quad x \in \mathbb{R}, \tag{2}$$

where $K(\cdot)$ is the continuous kernel function which is typically a bona-fide p.d.f. with zero mean and unit variance, $h = h(n) > 0$ is an arbitrary sequence of smoothing parameters (or bandwidths) that fulfills $\lim_{n\to\infty} h(n) = 0$, and $K_{x,h}(\cdot)$ will be the "continuous associated kernel" with the target x and the bandwidth h (cf. Section 1 for details in discrete case). Following the well-known expression (1) for unbounded supports of f, $K(\cdot)$ is classically symmetric and, therefore, the associated kernel is written as

$$K_{x,h}(\cdot) = \frac{1}{h}K\left(\frac{x - \cdot}{h}\right). \tag{3}$$

But the way from (2) to (1) is not always possible like for asymmetric associated kernels with respect to the target x. In the expressions (1) and (2), the bandwidth plays the role of a dispersion parameter around the target; this can be easily illustrated through the symmetric gaussian associated kernel $N_{x,h}$ with mean x (the target) and standard deviation h (the bandwidth) where $K = N_{0,1}$ (e.g. Senga Kiessé [27], pages 172–174). The expression (1) is known since Rosenblatt [22] and Parzen [21]. For recent references, one can see Tsybakov [31]. The works usually cited of Devroye [7], Scott [26] and Silverman [30] concern some generalities on (supposed) continuous data. For functional data, one can refere to Ferraty and Vieu [9]. The contributions of Simonoff [28] and Simonoff and Tutz [29] are concerned with ordered categorical and discrete data *always* using the continuous kernels. The second expression (2), that we will use in this paper, is from Chen [5,6] for adapting a "type of continuous kernel" generally asymmetric (such beta and gamma) to the support of f; see also Scaillet [24] for inverse Gaussian and reciprocal inverse Gaussian kernels. The case of a bounded support (from two or one end) of f to estimate induces a choice of type of asymmetric kernel, while the symmetric continuous kernels K does not have any important proper effects and can be used indifferently for smoothing functions on unbounded supports.

In order to estimate a probability mass function (p.m.f.) on $\mathbb{T}$ (*e.g.* $\mathbb{N}+p\mathbb{N}$ for $p \geq 0$, $\mathbb{Z}^d$, $\{0, 1, \ldots, N\}^d$, $d \in \mathbb{N}\backslash\{0\}$) using a discrete kernel method, the empirical or naive estimator is often used because of its good asymptotical properties. However, this Dirac type kernel estimator is not appropriate with small samples sizes. Furthermore, its great default is that it does not take into account observations around the target because its bandwidth is null or does not exist. Except the naive estimator,

Aitchison and Aitken [3] have been the pioneers of discrete kernel estimators in the sense of (2); see our Example 3. But the discrete kernel used has a unique shape and is appropriate for categorical data and finite discrete distributions; see also Li and Racine [18] and references therein. Thus the case of discrete kernels for count data is not investigated in the sense of our works since [17], except a first attempt of Marsh and Mukhopadhyay [20]. That attempt is only experimental and applied on univariate count data (i.e. $\mathbb{T} = \mathbb{N}$). A necessity of a discrete smoothing using discrete kernels out of the Dirac kernel is illustrated in Figure 3; for example, a binomial discrete kernel estimator is more interesting than the empirical estimator for a small sample size. Thus, we define and build a discrete associated kernel which asymptotically tends to the Dirac type kernel. It results in many applications of the discrete associated kernel method in literature such that nonparametric estimations of discrete weighted function (Kokonendji et al. [15]) and regression count function (Kokonendji et al. [16]). The present work is concentrated on providing some theoretical mathematical tools and asymptotical results for the discrete associated kernel estimator. In this way, it completes the papers cited previously which practically show the usefulness of the discrete associated kernel approach.

In this paper, we review more generally the estimator (2) with discrete associated kernels for a p.m.f. $f(x) = \Pr(X_i = x)$, for all x in $\mathbb{T}$. We point out some ingredients necessary to the construction and study of discrete kernel estimators. The rest of the paper is organized as follows. Section 2 presents the definition of a discrete associated kernel with some examples. In Section 3, the basic properties of the corresponding estimator $\widetilde{f}_n$ of f are given. The pointwise then global consistencies of $\widetilde{f}_n$ in quadratic mean are therefore proved. Section 4 illustrates some aspects of the estimators using standard discrete kernels of Poisson, binomial and negative binomial; the underdispersed ones will be more interesting for small samples sizes than the consistent estimators. Section 5 is devoted to some remarks on the bandwidth selection and the importance of the kernel choice. A simulation study and some possible extensions to continuous kernel in view of unified method are presented.

2 Discrete associated kernel

In order to simplify we assume that the support $\mathbb{T}$ of the p.m.f. f, to be estimated, is the count set $\mathbb{N}$. We then consider for $\mathbb{T} = \mathbb{N}$ the topology inherited from the standard one of the real line number $\mathbb{R}$. The easy and straightforward methodology of discrete associated kernels requires a clarification of the notions of integral, continuity and derivation of any discrete function f on $\mathbb{T}$.

Let us consider on $\mathbb{T}$ the counting measure $\mu = \sum_{y\in\mathbb{T}} \delta_y$, where δ_y denotes the Dirac mass at y. The integral on $\mathbb{T}_1 \subseteq \mathbb{T}$ is the following summation:

$$\int_{\mathbb{T}_1} f(x)\mu(dx) = \sum_{x\in\mathbb{T}_1} f(x).$$

The continuity of $f : \mathbb{T} \to \mathbb{R}$ at $x \in \mathbb{T} \subset \mathbb{R}$ can be defined as follows:

$$\forall\ \epsilon > 0,\ \exists\ \eta > 0 : \forall\ y \in]x-\eta, x+\eta[\cap \mathbb{T} \Rightarrow |f(y) - f(x)| < \epsilon. \tag{4}$$

Thus, we are not restrained to a particular form of continuation of discrete function. Moreover, we easily observe that any p.m.f. is bounded and continuous in the sense of (4). Note that, for $\eta > 0$ in (4), the notion of discrete neighbourhood $]x-\eta, x+\eta[\cap\mathbb{T}$ of x can be reduced to the single point $\{x\}$.

Finally, the finite difference of f on $\mathbb{T}$ is used instead of derivation on $\mathbb{R}$; see, for example, Schumaker ([25], page 343), Agarwal and Bohner [2] for other definitions. In the case $\mathbb{T} = \mathbb{N}$ of this work, we consider the finite difference $f^{(k)}(x)$ of order $k \in \mathbb{N} \setminus \{0\}$ at $x \in \mathbb{N}$ by the recursive relation:

$$f^{(k)}(x) = \left\{f^{(k-1)}(x)\right\}^{(1)} \text{ with } f^{(1)}(x) = \begin{cases} \{f(x+1) - f(x-1)\}/2 & \text{if } x \in \mathbb{N} \setminus \{0\} \\ f(1) - f(0) & \text{if } x = 0. \end{cases}$$

The $f^{(k)}(x)$ always exist and are some linear combinations of $f(x \pm j)$ for $j \in \{0, 1, \cdots, k\}$ and $x \pm j \in \mathbb{N}$. For instance, the particular case with $k = 2$ is given by

$$f^{(2)}(x) = \begin{cases} \{f(x+2) - 2f(x) + f(x-2)\}/4 & \text{if } x \in \mathbb{N} \setminus \{0, 1\} \\ \{f(3) - 3f(1) + 2f(0)\}/4 & \text{if } x = 1 \\ \{f(2) - 2f(1) + f(0)\}/2 & \text{if } x = 0. \end{cases} \tag{5}$$

We use these finite differences in the discrete Taylor expansion of $f(x)$ at any point $a \in \mathbb{T}$ such that

$$f(x) = \sum_{j=0}^{k} \frac{f^{(j)}(a)}{j!}(x-a)^j + o\{(x-a)^k\} \tag{6}$$

(see, for example, Schumaker [25], Theorem 8.61 - page 351). Note that if $a \notin \mathbb{T}$ then we have $f(a) = 0$. However, denoting by $\lfloor a \rfloor \in \mathbb{T}$ the nearest value (in the sense of usual topology of $\mathbb{R}$) to $a \in \mathbb{R} \setminus \mathbb{T}$ such that $a = \lfloor a \rfloor \pm \eta$ with $\eta > 0$, we extend the definition of $f(a)$ as

$$\begin{aligned} f(a) &= f(\lfloor a \rfloor \pm \eta) \\ &= f(\lfloor a \rfloor) \pm \epsilon, \quad \epsilon > 0 \\ &= f(\lfloor a \rfloor) \pm \eta f^{(1)}(\lfloor a \rfloor) + o(\eta). \end{aligned} \tag{7}$$

Thus, the expansion (6) of $f(x)$ can be done at any point $a \notin \mathbb{T}$ by using (7) of order 1. This is useful for studying a p.m.f. at its mean which would not belong in $\mathbb{T}$.

2.1 Definition

The discrete associated kernel introduced in (2) is defined as follows.

Definition 1 *Let $\mathbb{T}$ be the discrete support of the p.m.f. f, to be estimated, x a fixed target in $\mathbb{T}$ and $h > 0$ a bandwidth. A p.m.f. $K_{x,h}(\cdot)$ on support $\mathbb{S}_x$ (not depending on h) is said to be an associated kernel, if it satisfies the following conditions:*

$$x \in \mathbb{S}_x, \tag{8}$$

$$\lim_{h\to 0} \mathbb{E}(\mathcal{K}_{x,h}) = x, \tag{9}$$

$$\lim_{h\to 0} Var(\mathcal{K}_{x,h}) = 0, \tag{10}$$

where $\mathcal{K}_{x,h}$ is the discrete random variable whose p.m.f. is $K_{x,h}(\cdot)$.

Note that the condition "$h \to 0$ as $n \to \infty$" does not need to be included in Definition 1 because the sample size n is not yet required.

In order to construct a discrete associated kernel $K_{x,h}$ from a parametric discrete probability distribution K_θ, $\theta \in \Theta \subset \mathbb{R}^d$, on the support $\mathbb{S}_\theta$ such that $\mathbb{S}_\theta \cap \mathbb{T} \neq \emptyset$, we need to establish a correspondence between $(x, h) \in \mathbb{T} \times (0, \infty)$ and $\theta \in \Theta$. In what follows, we will call $K \equiv K_\theta$ the *type of discrete kernel* to make a difference from the classical notion of continuous kernel (1). In this context, the choice of the discrete associated kernel becomes important as well as that of the bandwidth. Moreover, we distinguish the discrete associated kernels said sometimes of "second order" of those said of "first order" which verify the two first conditions (8) and (9); see Section 4 below.

Remarks. (i) Given a type of discrete kernel K, the construction of any discrete associated kernel is obviously not unique.

(ii) The condition (8) can be replaced by $\bigcup_{x\in\mathbb{T}} \mathbb{S}_x \supseteq \mathbb{T}$. This implies that the discrete associated kernel takes into consideration the support $\mathbb{T}$ of the p.m.f. f, to be estimated. We sometimes have $\mathbb{S}_x = \mathbb{S}$ (not depending on x) and, therefore, $\mathbb{S}_x = \mathbb{T}$ for all x. If $\bigcup_{x\in\mathbb{T}} \mathbb{S}_x$ is not equal to $\mathbb{T}$ then we have a problem of boundary bias.

(iii) The condition (9) expresses that the information around the target is taken into account such that if $h \to 0$ then we find again the kernel of the naive estimator at the limit of the mean. This basical condition points out that the discrete associated kernel of the estimator (2) is a kind of variable kernel. It allows for more flexibility to construct different discrete associated kernel from any discrete distribution K; for example, $\mathbb{E}(\mathcal{K}_{x,h}) = x + h$ or $\mathbb{E}(\mathcal{K}_{x,h}) = x$. Thus, all the discrete associated kernels verifying (9) share the property that the shape of kernel changes according to the value of the target x where they are calculated. The variable shape does not concern only the point mass at x but at each $y \in \mathbb{S}_x$ such that the modal value stays in x as $h \to 0$; see Figure 1 for the shape of the binomial kernel presented in Section 4. The amount of smoothing obtained changes depending on the behaviour of the variance

$\mathrm{Var}(\mathcal{K}_{x,h})$ with respect to the target x.

(iv) The bandwidth parameter $h > 0$ allows to take into account the observations X_i near to the target $x \in \mathbb{T}$ (in the sense of stochastic distance of the discrete kernel K). The local dispersion $\mathrm{Var}(\mathcal{K}_{x,h})$ at each $x \in \mathbb{T}$ shows the importance of the discrete associated kernel $K_{x,h}$ chosen which imposes its variance property for a consistency.

(v) The behaviour desired in (10) allows the discrete associated kernel to tend to the kernel of the naive estimator which is the Dirac type kernel.

Note finally that, to specify the rate of convergence of the discrete random variable $\mathcal{K}_{x,h}$ to $x \in \mathbb{T}$, we may write the mean and the variance of $\mathcal{K}_{x,h}$ as

$$\mathbb{E}(\mathcal{K}_{x,h}) = x + hA(x;h), \text{ with } A(x;h) = o(1) \text{ (eventually uniformly)},$$

or

$$\mathbb{E}(\mathcal{K}_{x,h}) = x + h^{1/r}o(1), \text{ with } r \geq 1,$$

and

$$Var(\mathcal{K}_{x,h}) = hB(x;h) + O(h^2), \text{ with } B(x;h) \neq 0;$$

thus, the assumptions (8)-(10) are still satisfied. However, we do not investigate here these previous hypotheses and we concentrate on discrete associated kernels satisfying the general assumptions of Definition 1.

Figure 1 about here

2.2 Examples

We present below some examples of discrete kernels that fullfils conditions (8)-(10) for some counting, categorical or finite $\mathbb{T}$.

Example 1. The well-known discrete empirical (or naive) estimator might be viewed as a particular case of the discrete associated kernel (2) upon taking

$$K_{x,h}(y) = \mathbb{I}_{y=x} = \begin{cases} 1 \text{ if } y = x \\ 0 \text{ if } y \neq x, \end{cases} \quad \text{for any } x \in \mathbb{T} \text{ and any } h \geq 0, \tag{11}$$

where $\mathbb{I}_A$ denotes the indicator function of any given event A that takes the value 1 if the event A occurs and 0 otherwise. Note that the smoothing parameter h plays no role here and that the degenerate random variable $\mathcal{D}(x)$ associated to this discrete kernel satisfies (8), (9) and (10) with

$$\mathbb{S}_x = \{x\}, \quad \mathbb{E}\{\mathcal{D}(x)\} = x, \quad \mathrm{Var}\{\mathcal{D}(x)\} = 0.$$

Example 2. The following class of symmetric discrete kernels has been proposed

by Kokonendji et al. [17]. It generalizes the classical triangular kernel and might be constructed as follows. First, the support $\mathbb{T}$ of the p.m.f f, to be estimated, can be unbounded (*e.g.* $\mathbb{N}$, $\mathbb{Z}$) or finite (*e.g.* $\{0, 1, \cdots, N\}$). Then, suppose that h is a given bandwidth parameter and a is an arbitrary and fixed integer. For any x fixed in $\mathbb{T}$, consider the random variable $\mathcal{T}_{a;x,h}$ defined on $\mathbb{S}_x = \{x, x \pm 1, \cdots, x \pm a\}$ and whose p.m.f. is given by

$$K_{x,h}(y) = \frac{(a+1)^h - |y-x|^h}{P(a,h)}, \quad \forall y \in \mathbb{S}_x,$$

where $P(a,h) = (2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h$ is the normalizing constant. Since $K_{x,h}$ is symmetric around x, assumptions (8) and (9) are satisfied. As for the variance term (10), note that, for $a \in \mathbb{N}$ fixed, one has

$$\begin{aligned} V(a,h) &= \frac{1}{P(a,h)} \left\{ \frac{a(2a+1)(a+1)^{h+1}}{3} - 2\sum_{k=0}^{a} k^{h+2} \right\} \\ &\simeq \left\{ \frac{a(2a^2+3a+1)}{3} \log(a+1) - 2\sum_{k=1}^{a} k^2 \log(k) \right\} h + O(h^2), \end{aligned} \tag{12}$$

which does not depend on $x = \mathbb{E}(\mathcal{T}_{a;x,h})$ and tends to 0 when $h \to 0$. The last approximation holds for h sufficiently small. Note that a R package [1] for discrete triangular distributions is available. Recently, Kokonendji and Zocchi [13] have introduced a general version of discrete triangular distributions which helps for solving problems of boundary bias.

Example 3. Aitchison and Aitken [3] have introduced a discrete kernel estimator for categorical or finite discrete distribution (see also, Li and Racine [18]). Hence, we deduce its asymmetric discrete associated kernel that we present as follows. First, the support $\mathbb{T}$ of the p.m.f f, to be estimated, is finite with fixed size $c \in \mathbb{N} \setminus \{0, 1\}$. If the random variable X under investigation takes c different values, i.e. $\mathbb{T} := \{0, 1, \ldots, c-1\}$ (say), then, the discrete kernel in (2) might be

$$K_{x,h}(y) = (1-h)\mathbb{I}_{y=x} + \frac{h}{c-1}\mathbb{I}_{y \neq x}, \quad \forall y \in \mathbb{S}_x, \tag{13}$$

where h belongs to $(0, 1]$. In addition, the target x can be considered as the reference point of X and the smoothing parameter h is such that $1-h$ is the success probability of the reference point. Finally, if the bandwidth h goes to 0, then, the random variable $\mathcal{A}_{c;x,h}$ associated to $K_{x,h}$ will satisfy (8), (9), (10). Indeed, its support $\mathbb{S}_x = \mathbb{S}$ coincides with $\mathbb{T}$ and, its mean and variance are such that

$$\mathbb{E}(\mathcal{A}_{c;x,h}) = x + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right),$$

[1] http://cran.r-project.org/web/packages/TRIANG/index.html

$$\mathrm{Var}(\mathcal{A}_{c;x,h}) = -\left\{\frac{c^2(-2x+c-1)^2}{4(c-1)^2}\right\}h^2 + \left\{\frac{c(6x^2+2c^2-3c+1-6xc+6x)}{6(c-1)}\right\}h. \tag{14}$$

The distribution of $\mathcal{A}_{c;x,h}$ presents a uniform weight function on $\mathbb{S}_x \setminus \{x\}$ outside the reference point $x \in \mathbb{S}_x$. Some graphs of the distribution of $\mathcal{A}_{c;x,h}$ are presented in Senga Kiessé ([27], page 181).

Example 4. An extension of the discrete kernel in (13) to the integers set $\mathbb{Z}$ has been proposed by Wang and Van Ryzin [33]. More precisely, suppose that $\mathbb{T} = \mathbb{Z}$ and, for any x in $\mathbb{T}$ and h in $(0,1)$, denote by $\mathcal{Z}_{x,h}$ the random variable with support $\mathbb{S}_x = \mathbb{Z}$ and p.m.f.

$$K_{x,h}(z) = (1-h)\mathbb{I}_{z=x} + \frac{1}{2}(1-h)h^{|z-x|}\mathbb{I}_{|z-x|\geq 1}, \quad \forall z \in \mathbb{Z}.$$

Then, provided that $h \to 0$, this discrete kernel fulfills assumptions (8), (9) and (10) since

$$\mathbb{T} = \mathbb{S}_x, \quad \mathbb{E}(\mathcal{Z}_{x,h}) = x, \quad \text{and} \quad \mathrm{Var}(\mathcal{Z}_{x,h}) = h\frac{(1+h)}{(1-h)^2}.$$

3 Discrete associated kernel estimator

Let us give the first properties of the estimator (2) with a discrete associated kernel.

Proposition 1 *Let $X_1, X_2, \ldots, X_n$ be an n random sample i.i.d. from the unknown p.m.f. f on $\mathbb{T}$. Let $\widetilde{f}_n = \widetilde{f}_{n,h,K}$ be an estimator (2) of f with a discrete associated kernel. Then, for all $x \in \mathbb{T}$ and $h > 0$, we have*

$$\mathbb{E}\{\widetilde{f}_n(x)\} = \mathbb{E}\{f(\mathcal{K}_{x,h})\},$$

where $\mathcal{K}_{x,h}$ is the random variable associated to the p.m.f. $K_{x,h}$ on $\mathbb{S}_x$. Furthermore, we have $\widetilde{f}_n(x) \in [0,1]$ for all $x \in \mathbb{T}$, and

$$\sum_{x\in\mathbb{T}} \widetilde{f}_n(x) = C,$$

where $C = C(n;h,K)$ is a positive and finite constant if $\sum_{x\in\mathbb{T}} K_{x,h}(y) < \infty$ for all $y \in \mathbb{T}$.

Proof: First, for all $x \in \mathbb{T}$, we successively have

$$\mathbb{E}\{\widetilde{f}_n(x)\} = \sum_{y\in\mathbb{T}\cap\mathbb{S}_x} K_{x,h}(y)f(y) = \sum_{y\in\mathbb{T}\cap\mathbb{S}_x} f(y)\Pr(\mathcal{K}_{x,h} = y) = \mathbb{E}\{f(\mathcal{K}_{x,h})\},$$

because $f(y)K_{x,h}(y) = 0$ for $y \notin \mathbb{T}\cap\mathbb{S}_x$; thus, the formula is proved. Then, $\widetilde{f}_n(x) \in [0,1]$ proceeds immediately from $K_{x,h}(X_i) \in [0,1]$ for all X_i. Finally, by writing

$C = n^{-1} \sum_{i=1}^{n} \{\sum_{x \in \mathbb{T}} K_{x,h}(X_i)\}$ for all $h > 0$ and by noting that

$$\sum_{x \in \mathbb{T}} K_{x,h}(y) = \sum_{x \in \mathbb{T} \cap \mathbb{S}_x} K_{x,h}(y)$$

for all $y \in \mathbb{T} = \text{support}(X_i)$, it follows from it that:

- on the one hand $C > 0$ because $K_{x,h}(x) > 0$ for $y = x \in \mathbb{T} \cap \mathbb{S}_x$ and

- on the other hand $C < \infty$ because $0 \leq K_{x,h}(y) < 1$ for all $y \in \mathbb{T} \cap \mathbb{S}_x$ and $\sum_{x \in \mathbb{T}} K_{x,h}(y) < \infty$ for all $y \in \mathbb{T}$. □

Notice that $C = 1$ for the estimators (2) with discrete Aitchison-Aitken associated kernel (Example 3) and symmetric continuous associated kernels such Gaussian. In general we have $C \neq 1$ for the estimators (2), as with discrete triangular associated kernels (Example 2) and standard discrete kernels in Section 4. In practice, we calculate the constant C depending on observations before normalizing $\widetilde{f}_n$ to be a p.m.f. Without loss of generality, from now we assume $C = 1$.

In the following section, the asymptotic behaviour of the *mean squared error* (MSE) of $\widetilde{f}_n(x)$ is examined together with its strong consistency and asymptotic normality; see Abdous and Kokonendji [1] for details. Then the global consistency of $\widetilde{f}_n(x)$ in the sense of the *mean integrated squared error* (MISE) is investigated with some illustrations using discrete kernels of Examples 1–3 in Section 2.2. A detailed study of the asymptotic bias and variance of the discrete triangular associated kernel estimator is already provided in Kokonendji et al.[17].

3.1 Pointwise consistency and asymptotic normality

The first consistency result concerns the MSE of $\widetilde{f}_n(x)$.

Theorem 1 *Under assumptions (8)–(10), for any fixed x in $\mathbb{T}$, one has*

$$\lim_{n \to \infty} \mathbb{E}\{\widetilde{f}_n(x) - f(x)\}^2 = 0.$$

Proof: The usual bias-variance decomposition of MSE gives : $\mathbb{E}\{\widetilde{f}_n(x) - f(x)\}^2 = Var\{\widetilde{f}_n(x)\} + \left[\mathbb{E}\{\widetilde{f}_n(x)\} - f(x)\right]^2$. Next, the bias term satisfies

$$\begin{aligned} \mathbb{E}\{\widetilde{f}_n(x)\} - f(x) &= \sum_{y \in \mathbb{T} \cap \mathbb{S}_x} [f(y) - f(x)] \Pr(\mathcal{K}_{x,h} = y) - f(x) \sum_{y \in \overline{\mathbb{T}} \cap \mathbb{S}_x} \Pr(\mathcal{K}_{x,h} = y) \\ &=: B_1 + B_2, \end{aligned}$$

where $\overline{\mathbb{T}}$ denotes $\mathbb{R} \setminus \mathbb{T}$. In the first quantity B_1, the sum runs over y's belonging to $\mathbb{T} \cap \mathbb{S}_x$ and such that $y \neq x$. Thus, since both $\mathbb{T}$ and $\mathbb{S}_x$ are discrete sets, one

can find a finite constant $\eta_x > 0$ such that $|y - x| \geq \eta_x$ for any y in $\mathbb{T} \cap \mathbb{S}_x \setminus \{x\}$. Consequently, we can write

$$\begin{aligned}
|B_1| &\leq 2\Pr\Big(\mathcal{K}_{x,h} \in \mathbb{T} \cap \mathbb{S}_x \quad \text{and} \quad |\mathcal{K}_{x,h} - x| \geq \eta_x\Big) \\
&\leq 2\Pr\Big(|\mathcal{K}_{x,h} - x| \geq \eta_x\Big) \\
&\leq \frac{2}{\eta_x^2}\mathbb{E}\{(\mathcal{K}_{x,h} - x)^2\}.
\end{aligned}$$

The last inequality follows from the Tchebychev-Markov inequality. To conclude that B_1 goes to 0 as $n \to \infty$, note that $\mathbb{E}\{(\mathcal{K}_{x,h} - x)^2\} = Var(\mathcal{K}_{x,h}) + \{\mathbb{E}(\mathcal{K}_{x,h}) - x\}^2$ and use assumptions (9) and (10).

Next, to show that the second quantity B_2 converges to 0, remark that if $y \in \overline{\mathbb{T}} \cap \mathbb{S}_x$ then necessarily $y \neq x$ and consequently it satisfies $|y - x| \geq \eta_x$ for some finite $\eta_x > 0$. Similar arguments as those used above enable to conclude.

The convergence of the variance term stems from

$$Var\{\widetilde{f}_n(x)\} \leq \frac{1}{n}\sum_{y \in \mathbb{T} \cap S_x} f(y)\{\Pr(\mathcal{K}_{x,h} = y)\}^2 \leq \frac{1}{n}.$$

□

Remark. The assumptions under which the MSE of the discrete kernel estimator converges to zero, might, at first glance, appear striking. Indeed, we do not impose any apparent assumption on the bandwidth $h = h(n)$, but as we saw in the examples presented in Section 2, assumptions (9) and (10) hold provided that $\lim_{n\to\infty} h(n) = 0$. That said, we still do not have the usual assumptions encountered in kernel probability density function, i.e. $\lim_{n\to\infty}[h(n) + \{nh(n)\}^{-1}] = 0$.

Now, we simply state the strong consistency of $\widetilde{f}_n(x)$ and then the asymptotic normality of $\widetilde{f}_n(x)$; see Abdous and Kokonendji [1, Theorem 2.4] and [1, Theorem 2.5], respectively.

Theorem 2 *Under assumptions (8)–(10), for any fixed x in $\mathbb{T}$, we have*

$$\widetilde{f}_n(x) \xrightarrow{a.s.} f(x) \quad as \quad n \to \infty, \tag{15}$$

where $\xrightarrow{a.s.}$ denotes almost sure convergence.

Theorem 3 *Under assumptions (8)–(10), for any fixed x in $\mathbb{T}$ such that $f(x) > 0$, we have*

$$\frac{\widetilde{f}_n(x) - \mathbb{E}\{\widetilde{f}_n(x)\}}{\Big[Var\Big\{\widetilde{f}_n(x)\Big\}\Big]^{1/2}} \xrightarrow{d} \mathcal{N}(0,1) \quad as \quad n \to \infty, \tag{16}$$

where $\xrightarrow{d}$ denotes convergence in distribution and $\mathcal{N}(0,1)$ is the standard normal

distribution.

3.2 Global consistency and illustrations

The criterion to use for this consistency is the MISE of $\widetilde{f}_n = \widetilde{f}_{n,h,K}$ defined as

$$MISE = \sum_{x\in\mathbb{T}} \mathrm{Var}\{\widetilde{f}_n(x)\} + \sum_{x\in\mathbb{T}} \mathrm{bias}^2\{\widetilde{f}_n(x)\} = MISE(n,h,K,f). \tag{17}$$

Theorem 4 *Let f be a p.m.f on $\mathbb{T}$ with $\lim_{x\to\infty} f(x) = 0$. Then, the estimator (2) $\widetilde{f}_n = \widetilde{f}_{n,h,K}$ of f with any discrete associated kernel is such that, for $n \to \infty$ and $h = h(n) \to 0$, we have the behaviour*

$$\begin{aligned} MISE(n,h,K,f) = & \frac{1}{n}\sum_{x\in\mathbb{T}} f(x)[\{\Pr(\mathcal{K}_{x,h} = x)\}^2 - f(x)] \\ & + \sum_{x\in\mathbb{T}} \left[f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}\mathrm{Var}(\mathcal{K}_{x,h}) f^{(2)}(x)\right]^2 + o\left(\frac{1}{n} + h^2\right), \end{aligned}$$

where $f^{(2)}$ is the finite difference of second order given in (5).

Proof: Without loss of generality, we assume $\mathbb{S}_x \subseteq \mathbb{T}$ for all $x \in \mathbb{T}$. The pointwise variance of $\widetilde{f}_n$ can be written around the target x (which realizes the modal probability of $\mathcal{K}_{x,h}$) as:

$$\begin{aligned} \mathrm{Var}\{\widetilde{f}_n(x)\} &= \frac{1}{n}\left[\sum_{y\in\mathbb{S}_x} f(y)\left\{\Pr(\mathcal{K}_{x,h} = y)\right\}^2 - \left\{\sum_{y\in\mathbb{S}_x} f(y)\Pr(\mathcal{K}_{x,h} = y)\right\}^2\right] \\ &= \frac{1}{n} f(x)\{\Pr(\mathcal{K}_{x,h} = x)\}^2 - \frac{1}{n} f^2(x) + R_n(x;h), \end{aligned} \tag{18}$$

where the rest

$$\begin{aligned} R_n(x;h) = & \frac{1}{n}\sum_{y\in\mathbb{S}_x\setminus\{x\}} f(y)\left\{\Pr(\mathcal{K}_{x,h} = y)\right\}^2 + \frac{1}{n} f^2(x) \\ & - \frac{1}{n}\left[f(x) + \sum_{y\in\mathbb{S}_x}\{f(y) - f(x)\}\Pr(\mathcal{K}_{x,h} = y)\right]^2 \end{aligned}$$

is $o(1/n)$ under the hypothesis of discrete associated kernel; thus for all $x \in \mathbb{T}$, $R_n(x;h) \to 0$ when $n \to \infty$ and $h = h(n) \to 0$. Indeed, let $y \in \mathbb{S}_x \setminus \{x\}$ we can find a constant $\eta = \eta(y) > 0$ such that

$$
\begin{aligned}
0 &\leq \Pr(\mathcal{K}_{x,h} = y) \\
&\leq \Pr(|\mathcal{K}_{x,h} - x| > \eta) \\
&\leq \frac{1}{\eta^2}\mathbb{E}\{(\mathcal{K}_{x,h} - x)^2\} = \frac{1}{\eta^2}[\mathrm{Var}(\mathcal{K}_{x,h}) + \{\mathbb{E}(\mathcal{K}_{x,h}) - x\}^2] \to 0 \text{ when } h \to 0,
\end{aligned}
$$

and for $y = x$ we deduce the asymptotic modal probability $\Pr(\mathcal{K}_{x,h} = x) \to 1$ when $h \to 0$. Furthermore, we successively have:

$$
-f(x)\{\Pr(\mathcal{K}_{x,h} = x)\}^2 \leq 0,
$$

$$
-\frac{1}{n}\sum_{x\in\mathbb{T}} f(x)\{\Pr(\mathcal{K}_{x,h} = x)\}^2 + \frac{1}{n}\sum_{x\in\mathbb{T}} f^2(x) \leq \frac{1}{n}\sum_{x\in\mathbb{T}} f^2(x),
$$

$$
\sum_{x\in\mathbb{T}} R_n(x;h) \leq \frac{1}{n}\sum_{x\in\mathbb{T}} \mathrm{Var}\{\widetilde{f}_n(x)\} + \frac{1}{n}\sum_{x\in\mathbb{T}} f^2(x).
$$

Using $\mathrm{Var}\{\widetilde{f}_n(x)\} < \infty$ and $\lim_{n\to\infty} \mathrm{Var}\{\widetilde{f}_n(x)\} = 0$ with $0 < \sum_{x\in\mathbb{T}} f^2(x) < 1$ and $R_n(x;h) \geq 0$, we get $\sum_{x\in\mathbb{T}} R_n(x;h) \to 0$ when $n \to \infty$ and $h = h(n) \to 0$; hence, the result.

Concerning the pointwise bias of $\widetilde{f}_n$, we successively use the formula of Proposition 1 then the discrete Taylor expansion (6) of $f(\mathcal{K}_{x,h})$ at the point $\mathbb{E}(\mathcal{K}_{x,h})$ with (7) to obtain

$$
\begin{aligned}
\mathrm{bias}\{\widetilde{f}_n(x)\} &= \mathbb{E}\{f(\mathcal{K}_{x,h})\} - f(x) \\
&= f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}\mathrm{Var}(\mathcal{K}_{x,h}) f^{(2)}(x) + o(h). \qquad (19)
\end{aligned}
$$

Thus, the theorem ensues from (18) and (19) in the criterion of MISE (17). □

Application 1. In the very particular case of the Dirac type kernel (11) estimator and unbiased $\widetilde{f}_{n,0,D}$, the MISE (17) is equal to the integrated variance

$$
MISE(n,0,D,f) = \frac{1}{n}\sum_{x\in\mathbb{T}} f(x)\{1 - f(x)\} = \frac{1}{n}\left\{1 - \sum_{x\in\mathbb{T}} f^2(x)\right\}.
$$

This exact result is used as reference in comparison to the MISE of the others discrete associated kernel estimators, because $0 \leq \sum_{x\in\mathbb{T}} f^2(x) < 1$ and therefore we have the global consistency of the naive estimator as $MISE(n,0,D,f) \to 0$ when $n \to \infty$.

Application 2. For an estimator (2) with a discrete triangular associated kernel (cf. Example 2) with $a \in \mathbb{N}^*$ fixed, the exact $MISE(n,h,T_a,f)$ is obtained from

$$
\mathrm{bias}\{\widetilde{f}_n(x)\} = f(x)\left\{\frac{(a+1)^h}{P(a,h)} - 1\right\} + \sum_{y\in\mathbb{S}_x\setminus\{x\}} f(y)\Pr(\mathcal{T}_{a;x,h} = y)
$$

which tends to 0 when $h \to 0$, and

$$\mathrm{Var}\{\widetilde{f}_n(x)\} = \frac{1}{n}\left[f(x)\left\{\frac{(a+1)^h}{P(a,h)}\right\}^2 + \sum_{y\in\mathbb{S}_x\setminus\{x\}} f(y)\left\{\Pr(\mathcal{T}_{a;x,h}=y)\right\}^2\right] - \frac{1}{n}\left[f(x) + \sum_{y\in\mathbb{S}_x}\{f(y)-f(x)\}\Pr(\mathcal{T}_{a;x,h}=y)\right]^2$$

which tends to the variance $n^{-1}f(x)\{1-f(x)\}$ of the naive estimator when $h\to 0$. Theorem 4 leads to the expression of $MISE(n,h,T_a,f)$ as

$$MISE(n,h,T_a,f) = \frac{1}{n}\sum_{x\in\mathbb{T}} f(x)\left[\left\{\frac{(a+1)^h}{P(a,h)}\right\}^2 - f(x)\right] + \frac{1}{4}\{V(a,h)\}^2\sum_{x\in\mathbb{T}}\left\{f^{(2)}(x)\right\}^2 + o\left(\frac{1}{n}+h^2\right).$$

Thus, we have: $MISE(n,h,T_a,f) = o(1/n+h^2)\to 0$ when $n\to\infty$ and $h = h(n)\to 0$; because $\lim_{h\to 0}(a+1)^h/P(a,h) = 1$, $0\le\sum_{x\in\mathbb{T}} f(x)\{1-f(x)\} < 1$, $\lim_{h\to 0} V(a,h) = 0$ and $\sum_{x\in\mathbb{T}}\left\{f^{(2)}(x)\right\}^2 < \infty$ from (5). Hence, the estimators (2) with discrete triangular associated kernels are consistent in the sense of MISE.

Application 3. Concerning the estimator (2) with a discrete Aitchison-Aitken associated kernel (see Example 3) with $c\in\mathbb{N}\setminus\{0,1\}$ fixed, the exact $MISE(n,h,A_c,f)$ is obtained from

$$\mathrm{bias}\{\widetilde{f}_n(x)\} = \frac{-hc}{c-1}f(x) + \frac{h}{c-1}\sum_{i=0}^{c-1} f(i),$$

which also tends to 0 when $h\to 0$, and

$$\mathrm{Var}\{\widetilde{f}_n(x)\} = \frac{1}{n}\left[f(x)(1-h)^2 + \frac{h^2}{(c-1)^2}\left\{\sum_{i=0}^{c-1} f(i) - f(x)\right\}\right] - \frac{1}{n}\left[f(x)(1-h) + \frac{h}{c-1}\left\{\sum_{i=0}^{c-1} f(i) - f(x)\right\}\right]^2$$

which also tends to the variance $n^{-1}f(x)\{1-f(x)\}$ of the naive estimator when $h\to 0$. Using Theorem 4, the $MISE(n,h,A_c,f)$ is given by

$$MISE(n,h,A_c,f) = \frac{1}{n}\sum_{x\in\mathbb{T}} f(x)\{(1-h)^2 - f(x)\} + \sum_{x\in\mathbb{T}}\left[f\{\mathbb{E}(\mathcal{A}_{c;x,h})\} - f(x) + \frac{1}{2}\mathrm{Var}(\mathcal{A}_{c;x,h})f^{(2)}(x)\right]^2 + o\left(\frac{1}{n}+h^2\right).$$

Consequently, we have: $MISE(n,h,A_c,f) = o(1/n+h^2)\to 0$ as $n\to\infty$ and $h = h(n)\to 0$; because, for the variance part, $0\le\sum_{x\in\mathbb{T}} f(x)\{1-f(x)\} < 1$ and, for the bias part, $\lim_{h\to 0}\left[f\{\mathbb{E}(\mathcal{A}_{c;x,h})\} - f(x) + (1/2)\mathrm{Var}(\mathcal{A}_{c;x,h})f^{(2)}(x)\right] = 0$. Hence, the global consistency in quadratic mean of the estimator (2) with discrete

Aitchison-Aitken associated kernels is provided.

4 Other discrete kernels

In this section, we examine the case of the so-called s*tandard discrete kernels* which are discrete associated kernels of the first order (i.e. not verifying the condition (10) in Definition 1). They are built from usual discrete probability distributions of Poisson, binomial and negative binomial (see Johnson et al. [11]). They are also useful for smoothing a p.m.f f on $\mathbb{T} = \mathbb{N}$ or distributions of count data with small samples (see also Senga Kiessé [27], Chapter 1).

For all $x \in \mathbb{N}$ and $h > 0$, the discrete random variable $\mathcal{K}_{x,h}$ of standard discrete kernels satisfies, among other, the condition

$$\lim_{h \to 0} \mathrm{Var}(\mathcal{K}_{x,h}) \in \mathcal{V}(0) \tag{20}$$

which takes the place of (10) and where $\mathcal{V}(0)$ is a neighbourhood of 0 which does not depend on x. Here we present some standard discrete kernels such that $\mathbb{E}(\mathcal{K}_{x,h}) = x + h$; thus, the condition (9) holds for $\mathcal{K}_{x,h}$. In fact, it is more appropriate at the bound $x = 0$ and, in general, the target x is obviously not the mean of $\mathcal{K}_{x,h}$ which is asymmetric but rather its mode. From Theorem 4, the condition (20) does not allow the consistency in the sense of MISE of the corresponding estimators (2).

However, the standard discrete kernels estimators (2), *a fortiori* with small variances or *underdispersed* (variance $\leq$ mean), can be more interesting (in the sense of small MISE) for small samples sizes than the estimators (2) with discrete associated kernels or Dirac type kernel for which they have some good asymptotical properties. We give an element of proof by a graphic comparison of different exact MISE of a p.m.f. First, we present each of the three standard discrete kernels; then we give the bias and the variances of the corresponding estimators (2) which are sufficient to deduce their MISE using (17).

4.1 Poisson kernel

Consider a Poisson distribution $\mathcal{P}(\lambda)$ with $\lambda > 0$. For any x fixed in $\mathbb{T} = \mathbb{N}$ and $h > 0$, the corresponding random variable $\mathcal{P}_{x,h}$ associated to the Poisson kernel $P_{x,h}$ follows the distribution $\mathcal{P}(x + h)$ with support $\mathbb{S}_x = \mathbb{N}$ and p.m.f.

$$P_{x,h}(y) = \frac{(x+h)^y e^{-(x+h)}}{y!}, \quad \forall y \in \mathbb{N}.$$

Note that the discrete kernel proposed by Marsh and Mukhopadhyay (1999) inverts x and y in the expression $P_{x,h}(y)$ above and does not allow any mathematical study of properties. Our Poisson kernel $P_{x,h}$ is equidispersed (i.e. $\mathbb{E}(\mathcal{P}_{x,h}) = \mathrm{Var}(\mathcal{P}_{x,h}) = x+h$)

and has its mode between $x+h-1$ and $x+h$. Then, the discrete kernel $P_{x,h}$ fullfills assumptions (8) and (9) except (20). The corresponding estimator $\widetilde{f}_n$ has the pointwise bias

$$\text{bias}\{\widetilde{f}_n(x)\} = f(x)\left\{P_{x,h}(x) - 1\right\} + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)P_{x,h}(y)$$

which does not tend to 0 when $h \to 0$. Its pointwise variance can be written as

$$n\text{Var}\{\widetilde{f}_n(x)\} = f(x)P_{x,h}^2(x) + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)P_{x,h}^2(y) - \left[f(x) + \sum_{y\in\mathbb{N}}\{f(y)-f(x)\}P_{x,h}(y)\right]^2.$$

This $\widetilde{f}_n$ is not consistent in the sense of small MISE but can be more interesting than the naive estimator, for small or moderate samples sizes; see Figure 2 .

4.2 Binomial kernel

If we consider a binomial distribution $\mathcal{B}(N,p)$ with $N \in \mathbb{N}\setminus\{0\}$ and $p \in (0,1]$, we associate the random variable $\mathcal{B}_{x,h}$ corresponding to the binomial kernel $B_{x,h}$ following the distribution $\mathcal{B}\{x+1,(x+h)/(x+1)\}$ on $\mathbb{S}_x = \{0,1,\cdots,x+1\}$ for any $x \in \mathbb{N}$ and $h \in (0,1]$:

$$B_{x,h}(y) = \frac{(x+1)!}{y!(x+1-y)!}\left(\frac{x+h}{x+1}\right)^y\left(\frac{1-h}{x+1}\right)^{x+1-y}, \quad \forall y \in \mathbb{S}_x \subseteq \mathbb{N}.$$

It is an underdispersed discrete kernel (i.e. $\text{Var}(\mathcal{B}_{x,h}) = (x+h)(1-h)/(x+1)$ smaller than $\mathbb{E}(\mathcal{B}_{x,h}) = x+h$) having its mode around $x+h$. The binomial kernel satistfies the three assumptions (8), (9) and (20) with $\mathcal{V}(0) = [0,1)$. The bias and the variance of the corresponding estimator (2), for any $x \in \mathbb{N}$, are written as:

$$\text{bias}\{\widetilde{f}_n(x)\} = f(x)\left\{B_{x,h}(x) - 1\right\} + \sum_{y\in\mathbb{S}_x\setminus\{x\}} f(y)B_{x,h}(y)$$

and

$$n\text{Var}\{\widetilde{f}_n(x)\} = f(x)B_{x,h}^2(x) + \sum_{y\in\mathbb{S}_x\setminus\{x\}} f(y)B_{x,h}^2(y) - \left[f(x) + \sum_{y\in\mathbb{S}_x}\{f(y)-f(x)\}B_{x,h}(y)\right]^2.$$

The MISE of this estimator is not consistent but can be more smaller than those of estimators with discrete associated kernels and Dirac type kernel for some sample sizes not so large (see also Figure 2 below).

In the case of the negative binomial distribution $\mathcal{BN}(\lambda, p)$ with $\lambda > 0$ and $p > 0$, we define the binomial negative kernel $BN_{x,h}$ with the random variable $\mathcal{BN}_{x,h}$ following the distribution $\mathcal{BN}\{x+1, (x+1)/(2x+1+h)\}$ on $\mathbb{S}_x = \mathbb{N}$ for any $x \in \mathbb{N}$ and $h > 0$:

$$BN_{x,h}(y) = \frac{(x+y)!}{x!y!}\left(\frac{x+h}{2x+1+h}\right)^y\left(\frac{x+1}{2x+1+h}\right)^{x+1}, \quad \forall y \in \mathbb{N}.$$

It is an overdispersed discrete kernel (i.e. $\mathrm{Var}(\mathcal{BN}_{x,h}) = (x+h)\{1+(x+h)/(x+1)\}$ greater than $\mathbb{E}(\mathcal{BN}_{x,h}) = x+h$) having its mode around $x+h$. This dicrete kernel satisties assumptions (8), (9) but not (20). For an estimator (2) with a negative binomial kernel, the pointwise bias is given as

$$\mathrm{bias}\{\widetilde{f}_n(x)\} = f(x)\{BN_{x,h}(x) - 1\} + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)BN_{x,h}(y)$$

and the pointwise variance can be written as

$$n\mathrm{Var}\{\widetilde{f}_n(x)\} = f(x)BN^2_{x,h}(y) + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)BN^2_{x,h}(y) - \left[f(x) + \sum_{y\in\mathbb{N}}\{f(y)-f(x)\}BN_{x,h}(y)\right]^2.$$

Similarly to the previous cases, this estimator $\widetilde{f}_n$ is not consistent in the sense of small MISE; but, it can be more interesting than the naive estimator for some small samples sizes (see Figure 2 below).

Remark. For standard discrete kernel estimators, there exists a problem of boundary bias which can be solved as for continuous kernel estimators; see Zhang and Karunamuni [34] and also Chen [5,6]. Here, a solution is to use the Dirac type kernel at the bound $x = 0$ and a standard discrete kernel for any $x \in \mathbb{N} \setminus \{0\}$. For discrete triangular associated kernel estimators, this problem has arleady been solved in [17]; see also Kokonendji and Zocchi [13].

Comparisons of MISE: Figure 2 presents the comparative results of MISE of the estimators (2) with discrete kernels for the p.m.f.

$$f(x) = 0.4\,\frac{e^{-0.5}\,0.5^x}{x!} + 0.6\,\frac{e^{-10}\,10^x}{x!}, \quad \forall x \in \mathbb{N},$$

which is a mixture of two Poisson distributions with respective means $\mu_1 = 0.5$ and $\mu_2 = 10$. All computations were done by using the R software [23]. The intersection points between the MISE curves of each estimators (2) with discrete kernels and those of the MISE of the naive estimator point out the superior limit of n for which these estimators are more efficient than the naive. Beyond this limit, the naive estimator is better and its MISE tends to 0 as for the estimator (2) with a discrete triangular associated kernel under the condition that h is very small ($\to 0$). Among estimators (2) with standard discrete kernels, the binomial one seems consistent for h small (here 0.1) but it is only a visual effect. However, the binomial kernel

estimator is more interesting than the others here for small or moderate samples sizes and $h \leq 0.3$. The last observation is also illustrated through some discrete smoothing using appropriate bandwidth selection in Figure 3.

Figure 2 about here

5 Concluding remarks and extensions

5.1 Bandwidth selection

The bandwidth choice is generally realized in the sense of MISE (17) by approaching the ideal value of the bandwidth defined as

$$h_{id} = \arg\min_{h>0} MISE(n, h, K, f) = h_{id}(n, K, f).$$

Several methods already existing for continuous kernels can be adapted to the discrete case as the classical least-squares cross-validation method; see, for example, Bowman [4], Marron [19] and references therein. We simply propose two choices of bandwidth without making here a study on their consistencies.

Thus, for a given discrete kernel $K_{x,h}$ with $x \in \mathbb{T}$ and $h > 0$, we can prove that the optimal bandwidth h_{cv} of h is obtained by cross-validation as

$$h_{cv} = arg\min_{h>0} CV(h),$$

where

$$\begin{aligned} CV(h) &= \sum_{x\in\mathbb{T}} \{\widetilde{f}_n(x)\}^2 - \frac{2}{n}\sum_{i=1}^{n} \widetilde{f}_{n,-i}(X_i) \\ &= \sum_{x\in\mathbb{T}} \left\{\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\right\}^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} K_{X_i,h}(X_j) \end{aligned}$$

with $\widetilde{f}_{n,-i}(y) = (n-1)^{-1}\sum_{j\neq i} K_{y,h}(X_j)$ being computed as $\widetilde{f}_n(y)$ by excluding the observation X_i. This method is applied to all the estimators (2) with discrete kernels cited in this paper, independently on the support $\mathbb{T}$ of f to be estimated.

Let us consider the particular situation of count data ($\mathbb{T} = \mathbb{N}$) for which there exists a proportion of excess zeros $n_0 = \sharp(X_i = 0)$ in the sample $X_1, X_2, \ldots, X_n$. That corresponds to a well-known phenomenon (*e.g.* Kokonendji et al. [14], and references therein) for that we can use *the adapted bandwidth* $h_0 = h_0(n; K, f)$ of h as solution of

$$\sum_{i=1}^{n} \Pr(\mathcal{K}_{X_i,h_0} = 0) = n_0.$$

This last equation ensues from the expression $\mathbb{E}\{\widetilde{f}_n(x)\} = \sum_{y\in\mathbb{N}} f(y)\Pr(\mathcal{K}_{x,h} = y)$, where we take $y = 0$ and $f(0) = 1$ to identify the number of theoretical zeros into the empirical number of zeros n_0. According to the importance of zero-proportion in the sample, this bandwidth h_0 can be interesting (see Figure 3).

Table 1 about here

Finally, this selection method of h_0 by excess zeros does not apply to all discrete kernels. Indeed, h_0 exists for the three standard discrete kernels (see Table 1: explicit for Poisson and implicit for the other ones); but h_0 does not exist for discrete triangular associated kernels (see Kokonendji et al. [17] and Senga Kiessé [27], pages 84-85) and the method is not applicable to discrete Aitchison-Aitken associated kernel for categorical data.

Figure 3 about here

5.2 Simulation study

We consider simulated data from a Poisson distribution with mean $\mu = 2$. In order to measure the performance of estimators, we use the practical

$$ISE = \sum_{x\in\mathbb{N}} \left\{\widetilde{f}_n(x) - f(x)\right\}^2 = ISE(n, h, K, f)$$

and the theoretical MISE in (17). In Table 2, we calculate the optimal average of ISE and their standard errors for the estimators based on 1000 replications. For each simulation, the optimal discrete smoothing bandwidths are given by the cross-validation method. The optimal ISE are determined by using the optimal bandwidths. In general, for small or moderate samples sizes, the results in Table 2 show that discrete kernel estimators are more interesting that the empirical estimator. In addition, the simulated sum of the rest R_n in (18) is calculated in Table 3. On the one hand, we have $\sum_{x\in\mathbb{N}} R_n(x; h) \to 0$ as $n \to \infty$ for discrete triangular associated kernels with $a = 1$ and $a = 2$; on the other hand, the binomial kernel has a sum $\sum_{x\in\mathbb{N}} R_n(x; h)$ smaller than those of the two others standard discrete kernels.

Tables 2 and 3 about here

5.3 About discrete kernel choice

Similarly, the ideal discrete kernel satisfies

$$K_{id} = \arg\min_K MISE(n, h, K, f) = K_{id}(n, h, f).$$

Since the discrete (associated) kernel $K_{x,h}$ depends on the support $\mathbb{T}$ of f and also on each target $x \in \mathbb{T}$, we have to restrict us to a specifical class of discrete kernels

for realizing the optimization.

Thus, without loss of generality, we consider two random variables $\mathcal{K}_{x,h}^{[1]}$ and $\mathcal{K}_{x,h}^{[2]}$ connecting to discrete associated kernels (of first or second order) $K_{x,h}^{[1]}$ and $K_{x,h}^{[2]}$ on comparable supports $\mathbb{S}_x^{[1]}$ and $\mathbb{S}_x^{[2]}$, respectively. Up to

$$\mathbb{E}\left(\mathcal{K}_{x,h}^{[1]}\right) = \mathbb{E}\left(\mathcal{K}_{x,h}^{[2]}\right), \quad \forall x \in \mathbb{T} \text{ and } h > 0, \tag{21}$$

the discrete kernel $K^{[1]}$ is said to be better than the discrete kernel $K^{[2]}$ if and only if

$$\text{Var}\left(\mathcal{K}_{x,h}^{[1]}\right) \leq \text{Var}\left(\mathcal{K}_{x,h}^{[2]}\right), \quad \forall x \in \mathbb{T} \text{ and } h > 0.$$

The ingredients of this criterion of efficiency are partially presented in the Theorem 4 for n large and h small.

For example, the better discrete kernel in the family of triangular T_a with $a \in \mathbb{N}^*$ is obtained for $a = 1$. As for the class of standard discrete kernels, the better discrete kernel is the binomial one which is underdispersed. However, it is not possible with this indicator to compare systematically these two discrete kernel families; because of (21), we should take into consideration n, h and f, and therefore the MISE (see, for example, Figures 2, 3 and Table 2).

After all, for $h > 0$ and a p.m.f f, the choice of a discrete kernel depends on the sample size n. For n large, the Dirac type kernel or a discrete associated kernel will be sufficient to get a good discrete smoothing. However, an other discrete kernel which is not associated kernel but satisfying (20) can be also efficient. Concerning small samples sizes for which the Dirac type kernel is not appropriate, the use of a discrete kernel of first order verifying (20) or a discrete associated kernel is more interesting.

Table 4 about here

5.4 Extensions

The estimator (2) opens another way for nonparametric approach by continuous or discrete associated kernel and therefore mixed, because Definition 1 concerning the discrete associated kernels can be extended. Indeed, the symmetric continuous kernels (3) and the asymmetric ones in Tables 5 and 6, respectively, fullfill assumptions (8)-(10) of associated kernels; see Senga Kiessé [27], pages 172-178. Thus, we give the extension of Definition 1 for continuous associated kernels as follows.

Definition 2 *Let $\mathbb{T}$ be a continuous support of f, to be estimated, x a fixed target in $\mathbb{T}$ and $h > 0$ a bandwidth. A p.d.f. $K_{x,h}(\cdot)$ on support $\mathbb{S}_{x,h}$ is said to be an associated kernel, if it satisfies assumptions (8)–(10) with*

$$\mathbb{E}(\mathcal{K}_{x,h}) = \int_{\mathbb{S}_{x,h} \cap \mathbb{T}} y K_{x,h}(y) dy, \tag{22}$$

$$Var(\mathcal{K}_{x,h}) = \int_{\mathbb{S}_{x,h}\cap\mathbb{T}} \{y - \mathbb{E}(\mathcal{K}_{x,h})\}^2 K_{x,h}(y)dy, \tag{23}$$

where $\mathcal{K}_{x,h}$ is the continuous random variable whose p.d.f. is $K_{x,h}(\cdot)$.

The following result provides an interpretation for the well-known symmetric continuous kernels (3); see, for example, Table 5.

Proposition 2 *If $K(\cdot)$ is a symmetric continuous kernel function which is a bona-fide p.d.f. with zero mean and unit variance. Then, for x fixed in $\mathbb{T}$ and $h > 0$, the associated kernel $(1/h)K\{(x - \cdot)/h\} = K_{x,h}(\cdot)$ on support $\mathbb{S}_{x,h}$ is also symmetric with mean x and standard deviation h.*

Proof: Without loss of generality, we assume that $\mathbb{S}_{x,h} \subseteq \mathbb{T}$. By writing $t = (y-x)/h$ then $dy = hdt$ in (22) and (23), it is easy to check the results using $K(-t) = K(t)$, $\int_{\mathbb{S}_{x,h}} tK(t)dt = 0$ and $\int_{\mathbb{S}_{x,h}} t^2K(t)dt = 1$. □

Tables 5 and 6 about here

Some properties which are avalaible in discrete case can be applied to continuous case, like the choice of the kernel depending on the support $\mathbb{T}$ of the function to be estimated and also the consistencies. The first extensions of this work will be to build an associated kernel estimator on time scales $\mathbb{T}$ (e.g. Agarwal and Bohner [2]), then to detail in this case some notions as integral, continuity, derivation of a function and also boundary bias (e.g. Chen [5,6], Zhang and Karunamuni [34]). A basic of mathematical tools is given for discrete kernels summarized in Table 4; for example, the discrete Taylor expansion can be improved in order to unify the theory. Several points are still to be proved such that the good behaviour of the binomial kernel for small samples sizes, other optimal choices of discrete smoothing bandwidth and the crucial problem of an optimal associated kernel (discrete or continuous). New discrete probability distributions may also be constructed, as the class of discrete triangular (e.g. [13]), for serving as discrete associated kernels.

Finally, the discrete associated kernels methodology is useful in various domains as in actuarial or demography. Moreover, two others extensions of this associated kernels method that will be of interest are to consider the multidimentional case in the estimation of a discrete or mixed regression function and the nonparametric weighted Poisson regression problems. Some works in these direction are in progress with respect to the discrete associated kernel approach.

Acknowledgements. We thank the anonymous referee and Associate Editor for their valuable comments. We are grateful to Belkacem Abdous for several discussions and his careful reading.

References

[1] B. Abdous, C.C. Kokonendji, Consistency and asymptotic normality for discrete associated kernel estimator, African Diaspora Journal of Mathematics 8 (2009) 63–70.

[2] R.P. Agarwal, M. Bohner, Basic calculus on time scales and some of its applications, Results in Mathematics 35 (1999) 3–22.

[3] J. Aitchison, C.G.G. Aitken, Multivariate binary discrimination by the kernel method, Biometrika 63 (1976) 413–420.

[4] A. Bowman, An alternative method of cross-validation for the smoothing of density estimates, Biometrika 71 (1984) 352–360.

[5] S.X. Chen, Beta kernels estimators for density functions, Computational Statistics & Data Analysis 31 (1999) 131–145.

[6] S.X. Chen, Gamma kernel estimators for density functions, Annals of the Institute of Statistical Mathematics 52 (2000) 471–480.

[7] L. Devroye, A Course in Density Estimation, Birkhäuser, Boston, 1987.

[8] W. Feller, An Introduction to Probability Theory and Its Applications, 2nd ed., John Wiley & Sons, New York, 1966.

[9] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer, Berlin, 2006.

[10] E. Hille, Functional Analysis and Semi-Groups, American Statistical Society Colloquium, New York, 1948.

[11] N.L. Johnson, A.W. Kemp, S. Kotz, Univariate Discrete Distributions, 3rd ed., John Wiley & Sons, Hoboken, New Jersey, 2005.

[12] M.C. Jones, Estimating densities, quantiles, quantile densities and density quantiles, Annals of the Institute of Statistical Mathematics 44 (1992) 721–727.

[13] C.C. Kokonendji, S.S. Zocchi, Extensions of discrete triangular distributions and boundary bias in kernel estimation for discrete functions, Statistics & Probability Letters 80 (2010) 1655–1662.

[14] C.C. Kokonendji, D. Mizère, N. Balakrishnan, Connections of the Poisson weight function to overdispersion and underdispersion, Journal of Statistical Planning and Inference 138 (2008), 1287–1296.

[15] C.C. Kokonendji, T. Senga Kiessé, N. Balakrishnan, Semiparametric estimation for count data through weighted distributions, Journal of Statistical Planning Inference 139 (2009), 3625–3638.

[16] C.C. Kokonendji, T. Senga Kiessé, C.G.B. Demétrio, Appropriate kernel regression on a count explanatory variable and applications, Advances and Applications in Statistics, 12 (2009), 99–126.

[17] C.C. Kokonendji, T. Senga Kiessé, S.S. Zocchi, Discrete triangular distributions and non-parametric estimation for probability mass function, Journal of Nonparametric Statistics 19 (2007) 241–254.

[18] Q. Li, J.S. Racine, Nonparametric Econometrics: Theory and Practice, Princeton University Press, Princeton and Oxford, 2007.

[19] J.S. Marron, A comparison of cross-validation techniques in density estimation, The Annals of Statistics 15 (1987) 152–162.

[20] L.C. Marsh, K. Mukhopadhyay, Discrete Poisson kernel density estimation with an application to wildcat coal strikes. Applied Economics Letters 6 (1999) 393–396.

[21] E. Parzen, On estimation of a probability density function and mode, Annals of Mathematical Statistics 33 (1962) 1065–1076.

[22] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Annals of Mathematical Statistics 27 (1956) 832–837.

[23] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, 2008. URL http://www.R-project.org.

[24] O. Scaillet, Density estimation using inverse and reciprocal inverse Gaussian kernels, Journal of Nonparametric Statistics 16 (2004) 217–226.

[25] L.L. Schumaker, Spline Functions: Basic Theory, Wiley, New York, 1981.

[26] D.W. Scott, Multivariate Density Estimation - Theory, Practice, and Visualization, Wiley, New York, 1992.

[27] T. Senga Kiessé, Nonparametric Approach by Discrete Associated-Kernel for Count Data, Ph.D. manuscript (in French), University of Pau, 2009. URL http://tel.archives-ouvertes.fr/tel-00372180/fr/

[28] J.S. Simonoff, Smoothing Methods in Statistics, Springer, New York, 1996.

[29] J.S. Simonoff, G. Tutz, Smoothing methods for discrete data. In: Smoothing and Regression: Approaches, Computation, and Application, ed. M.G. Schimek, Wiley, New York, 2000, pp. 193–228.

[30] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986.

[31] A.B. Tsybakov, Introduction à l'Estimation Non-Paramétrique, Springer, Paris, 2004.

[32] M.P. Wand, M.C. Jones, Kernel smoothing, Chapman & Hall, London, 1995.

[33] M.-C. Wang, J. Van Ryzin, A class of smooth estimators for discrete distributions, Biometrika 68 (1981) 301–309.

[34] S. Zhang, R.J. Karunamuni, Some improvements on a boundary corrected kernel density estimator, Statistics & Probability Letters 78 (2008) 499–507.

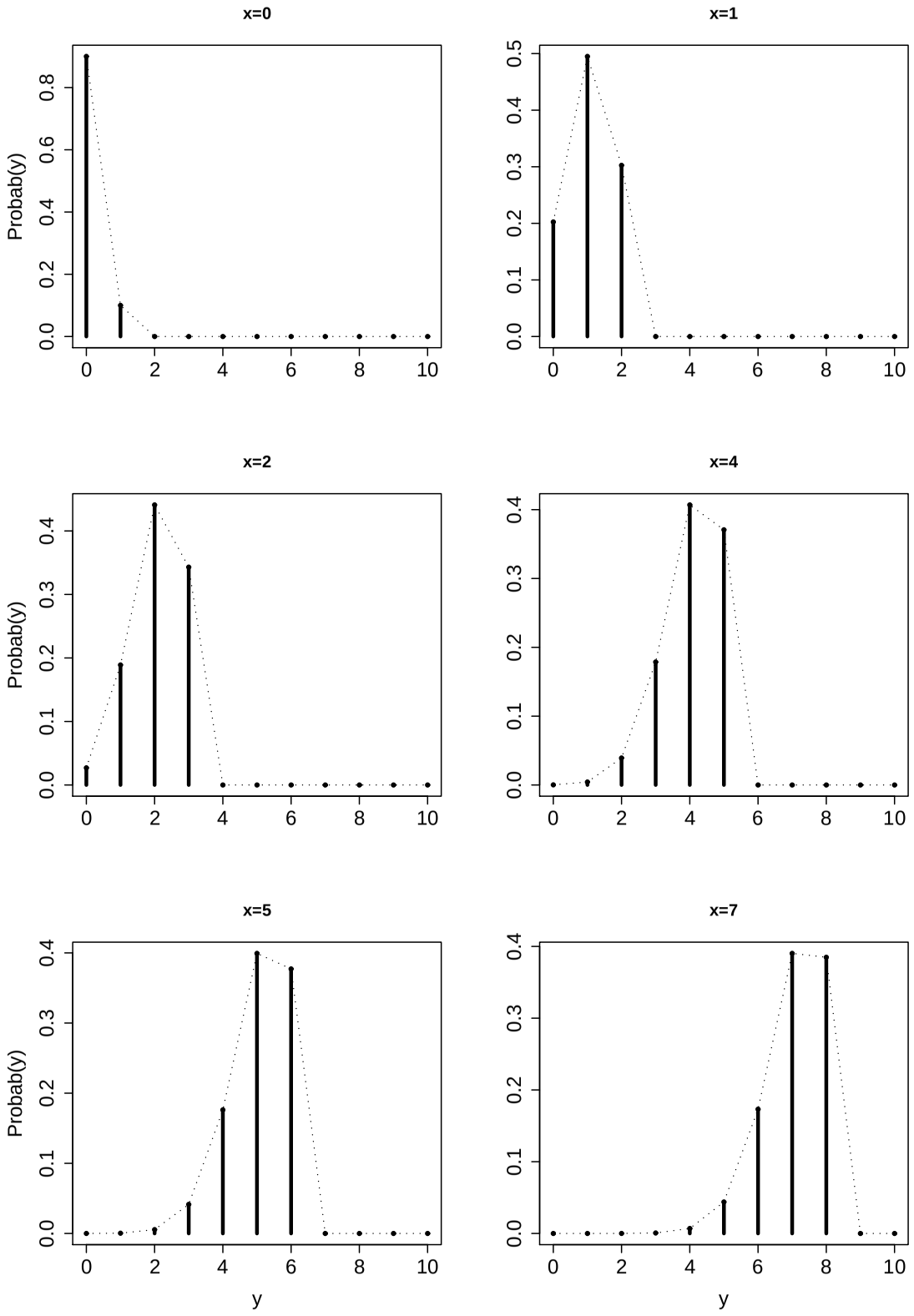

Fig. 1. Shape of the binomial kernel for $h = 0.1$ and some values of x

Table 1
Solutions h_0 for standard discrete kernels

Type of kernel	h_0 such that $\sum_{i=1}^{n} \Pr\left(\mathcal{K}_{X_i,h_0} = 0\right) = n_0$
Poisson	$h_0 = \log\left(\frac{1}{n_0}\sum_{i=1}^{n} e^{-X_i}\right)$
Binomial	$\sum_{i=1}^{n}\left(\frac{1-h_0}{X_i+1}\right)^{X_i+1} = n_0$
Negative binomial	$\sum_{i=1}^{n}\left(\frac{X_i+1}{2X_i+1+h_0}\right)^{X_i+1} = n_0$

Table 2
Simulated MISE and optimal average ISE and their standard errors (in parentheses) for discrete kernel and empirical estimators. The results multiplied by 10^3 are given

Kernel	Triangular $a=1$		Triangular $a=2$		Binomial	
n	$MISE$	$\mathbb{E}(ISE)$	$MISE$	$\mathbb{E}(ISE)$	$MISE$	$\mathbb{E}(ISE)$
20	10.82	14.10 (15.94)	15.37	14.28 (13.14)	12.81	21.52 (23.36)
50	5.47	6.81 (7.07)	9.89	8.63 (5.91)	7.13	7.10 (7.87)
80	4.02	4.80 (4.37)	5.60	6.34 (4.23)	4.92	4.46 (3.68)
100	3.63	4.14 (3.55)	4.95	5.42 (3.43)	4.60	3.94 (3.41)
200	2.64	2.57 (1.83)	3.94	2.82 (1.98)	3.47	2.72 (2.01)
500	1.22	1.28 (0.87)	1.57	1.46 (0.94)	2.86	2.03 (1.12)
700	0.93	0.93 (0.62)	1.12	1.17 (0.71)	2.75	1.91 (0.92)
Kernel	Poisson		Negative binomial		Empirical	
n	$MISE$	$\mathbb{E}(ISE)$	$MISE$	$\mathbb{E}(ISE)$	$MISE$	$\mathbb{E}(ISE)$
20	21.15	17.88 (9.42)	27.83	27.60 (7.58)	39.65	36.48 (24.73)
50	15.73	15.76 (5.76)	26.76	27.73 (5.39)	15.86	15.84 (10.76)
80	14.68	15.02 (4.09)	26.34	27.63 (4.31)	9.91	10.03 (6.97)
100	13.80	15.07 (3.68)	24.85	27.78 (4.07)	7.93	8.13 (5.97)
200	14.04	14.95 (2.50)	25.80	28.26 (2.99)	3.96	3.89 (2.76)
500	13.62	13.75 (1.59)	25.00	26.76 (2.09)	1.59	1.57 (1.09)
700	13.58	12.95 (1.07)	24.94	25.96 (1.87)	1.13	1.14 (0.78)

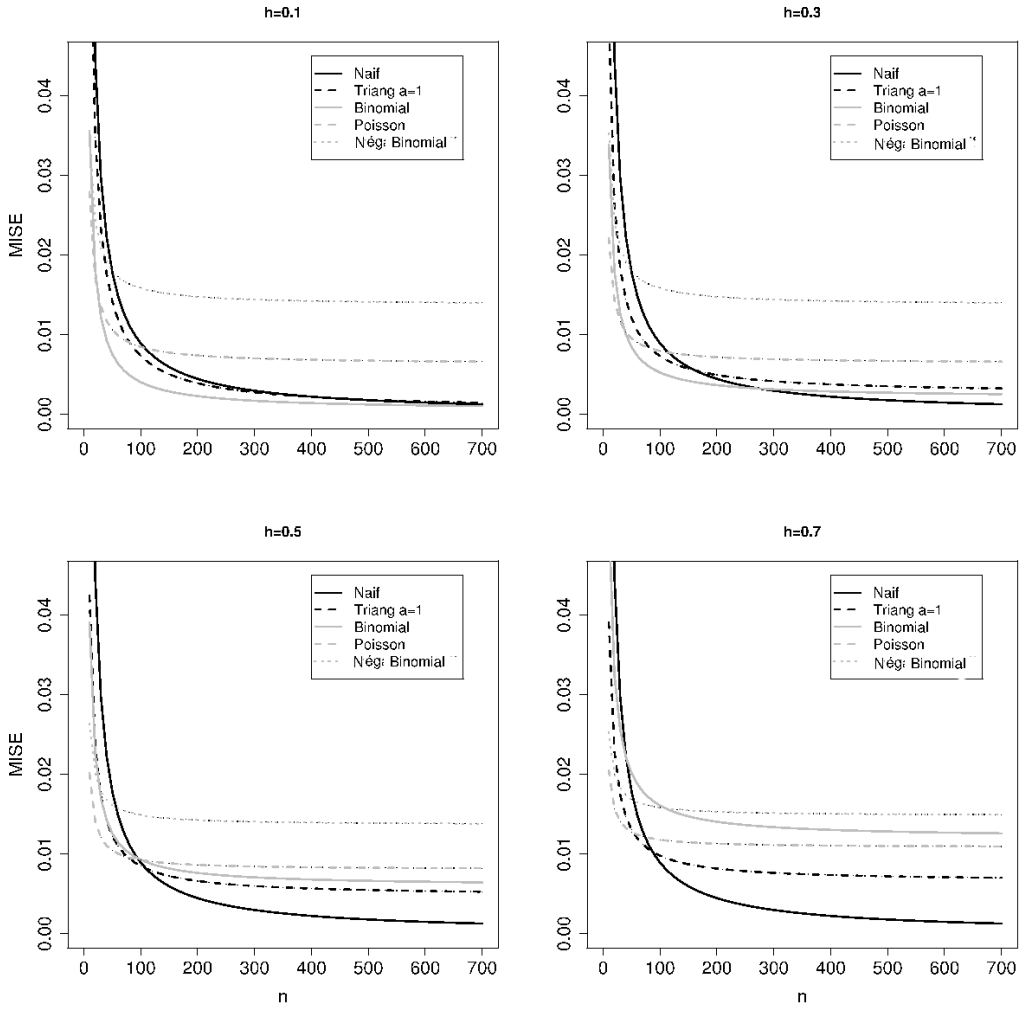

Fig. 2. Comparison of the exact MISE of discrete kernel estimators for $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

Table 3
Simulated $\sum_{x\in\mathbb{N}} R_n(x;h)$ of discrete kernel estimators for $f = \mathcal{P}(2)$. The results multiplied by 10^3 are given

Kernel	Triang. $a=1$	Triang. $a=2$	Binomial	Poisson	Negative binomial
n	$\sum_{x\in\mathbb{N}} R_n(x;h)$				
50	3.95	3.78	2.71	3.11	5.65
200	0.19	0.13	0.67	0.77	1.53
500	0.057	0.016	0.26	0.31	0.61
700	0.035	0.013	0.18	0.22	0.44

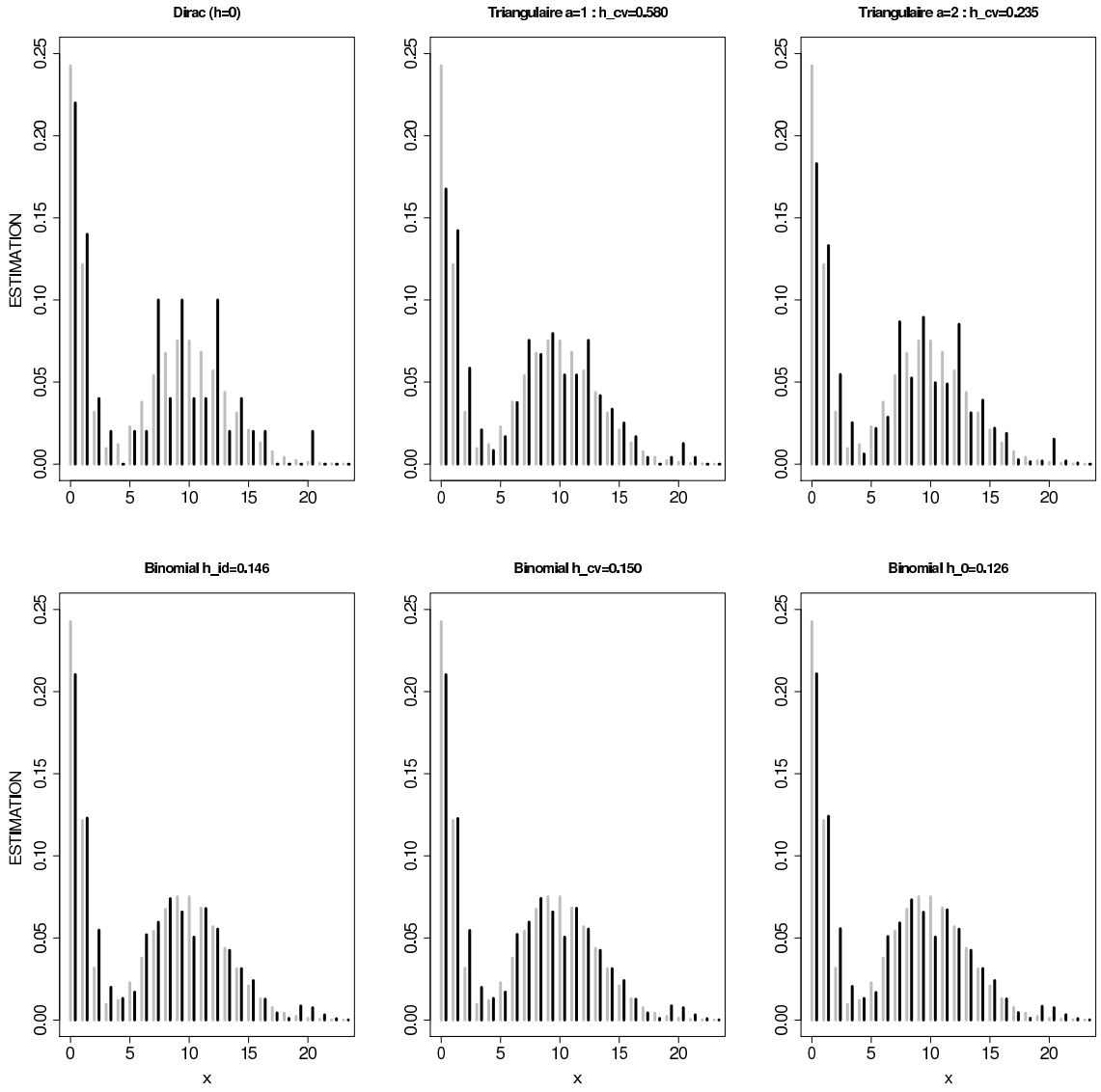

Fig. 3. Discrete smoothing (black line) by discrete kernel estimators for simulated data of $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$ (grey line) with $n = 50$

Table 4
Summary of some properties of discrete kernel estimators

Type of discrete kernel	$\mathbb{E}(\mathcal{K}_{x,h})$	$\mathrm{Var}(\mathcal{K}_{x,h})$	$\lim_{h\to 0}\mathrm{Var}(\mathcal{K}_{x,h})$	Convergence of MISE	Cross-validation	Excess of zero	Symmetry of $\mathcal{K}_{x,h}$	Remarks
Dirac	x	0	0	YES ($n \nearrow \infty$)	$--$	$--$	YES	No bandwidth
Poisson	$x+h$	$x+h$	$x \in \mathbb{N}$	NO	YES	YES	NO	Equi-dispersion
Binomial	$x+h$	$(x+h)\left(\frac{1-h}{x+1}\right)$	$0 \leq \frac{x}{x+1} < 1$	NO	YES	YES	NO	Under-dispersion
Negative binomial	$x+h$	$(x+h)\left(1+\frac{x+h}{x+1}\right)$	$\frac{x(2x+1)}{x+1} \geq 0$	NO	YES	YES	NO	Over-dispersion
Triangular $a \in \mathbb{N}\setminus\{0\}$	x	$V(a,h)$: see (12)	0	YES ($n \nearrow \infty$ and $h \searrow 0$)	YES	NO	YES	Boundary bias
Aitchison-Aitken	$x+o(h)$	$\mathrm{Var}(\mathcal{A}_{c;x,h})$: see (14)	0	YES ($n \nearrow \infty$ and $h \searrow 0$)	YES	$--$	NO	Categorical data
Wang-Van Ryzin	x	$\frac{h(h+1)}{(1-h)^2}$	0	YES ($n \nearrow \infty$ and $h \searrow 0$)	YES	$--$	YES	Boundary bias

Table 5
Examples of classical symmetric continuous kernels (Wand and Jones [32], page 31)

Kernel	Density	Support	Efficiency
Epanechnikov	$(3/4)(1-u^2)\mathbb{I}_{[-1,1]}$	$[-1,1]$	1.000
Biweight	$(15/16)(1-u^2)^2\mathbb{I}_{[-1,1]}$	$[-1,1]$	0.994
Triangular	$(1-\lvert u\rvert)\mathbb{I}_{[-1,1]}$	$[-1,1]$	0.986
Gaussian	$(1/\sqrt{2\pi})\exp(-u^2/2)$	$\mathbb{R}$	0.951
Uniform	$(1/2)\mathbb{I}_{[-1,1]}(u)$	$[-1,1]$	0.930

Table 6
Summary of some properties of asymmetric continuous associated kernel

Kernel	Support	Associated kernel	$\mathbb{E}(\mathcal{K}_{x,h})$	$\mathrm{Var}(\mathcal{K}_{x,h})$
Beta(a,b) [5]	$[0,1]$	$a = xh^{-1}+1$	$\frac{x+h}{2h+1}$	$\frac{x(1-x)h+h^2+h^3}{(1+2h)^2(1+3h)}$
		and $b=(1-x)h^{-1}+1$		
Ga(a,b) [6]	$[0,+\infty)$	$a = xh^{-1}+1$	$x+h$	$xh+h^2$
		and $b=h$		
IG(a,b) [24]	$(0,+\infty)$	$a=x$	x	x^3h
		and $b=h^{-1}$		
RIG(a,b) [24]	$(0,+\infty)$	$a=(x-h)^{-1}$	x	$xh+h^2$
		and $b=h^{-1}$		